Transductive and Matched-Pair Machine Learning for Difficult Target Detection Problems

James Theiler

Los Alamos National Laboratory, Los Alamos, NM 87545

ABSTRACT

This paper will describe the application of two non-traditional kinds of machine learning (transductive machine learning and the more recently proposed matched-pair machine learning) to the target detection problem. The approach combines explicit domain knowledge to model the target signal with a more agnostic machine-learning approach to characterize the background. The concept is illustrated with simulated data from an elliptically-contoured background distribution, on which a subpixel target of known spectral signature but unknown spatial extent has been implanted.

Keywords: machine learning, target detection, transductive, matched-pair

1. INTRODUCTION

Machine learning (ML) algorithms are potentially very powerful because they are flexible enough to model arbitrary distributions yet rigorous enough that one can prove theorems about their performance.[2] The use of traditional ML for classification of remote sensing imagery is relatively straightforward and has already met with considerable success.[3–6] The target detection problem is superficially similar to classification (it aims to distinguish target from non-target) but is in several important ways quite different. For target detection, only a few on-target training samples are typically available, the target strength can be variable, and the operation typically must be performed in a very low false alarm rate regime. Thus, although many of the tools of ML – kernels, regularization, cross-validation, *etc.*– have proven useful for a number of remote sensing tasks, the traditional formulation of ML is not well suited to the target detection problem.

The discussion begins in Section 2 with a description of three traditional machine learning scenarios: supervised, unsupervised, and semi-supervised. This will be followed by two extensions of traditional machine learning – transductive and matched-pair – which work together to enhance the ability of machine learning to be used for target detection problems.

One thing the three traditional scenarios have in common is a rigorous distinction between in-sample and out-of-sample data, between training and exploitation. Transductive learning, described in Section 3, provides new scenarios where this distinction is blurred, but in a way that maintains the rigor, and can lead to better performance (a "transductive advantage").

Another emphasis in traditional machine learning is on the utility of discriminative (versus generative) models; such models are less useful for describing a physical reality, but often more effective at identifying the boundary between distinct classes from a limited number of data samples. Section 4 introduces matched-pair machine learning, which provides a way to incorporate physics modeling of the target with more generic machine learning tools for target detection problems. Although matched-pairs, as originally introduced,[1] employ a binary target/non-target dichotomy, many practical target detection problems include a nuisance parameter, for instance characterizing target size or strength, and the matched-pair scheme needs to be modified to account for this ambiguity.

2. TRADITIONAL MACHINE LEARNING

In supervised binary classification – *the* classic machine learning problem – one has a finite set of labeled samples: $(\mathbf{x}_1, y_i), \ldots, (\mathbf{x}_n, y_n)$ where $\mathbf{x} \in \mathbb{R}^d$ is a d-dimensional vector of attributes and $y \in \{-1, +1\}$ is a label. The goal is to infer a function $f(\mathbf{x})$ that predicts the label y that is associated with data vectors $\mathbf{x}$ whose labels are *not* known. We assume that the samples $(\mathbf{x}_i, y_i)$ were drawn from an underlying distribution $p(\mathbf{x}, y)$; we assume this distribution *exists*, but we don't assume we know what it is. We furthermore assume that the out-of-sample data points are drawn from the same distribution, and although we know the attribute vectors $\mathbf{x}$ for those out-of-sample points, we do not know their labels. That's what we want machine learning to help us with.

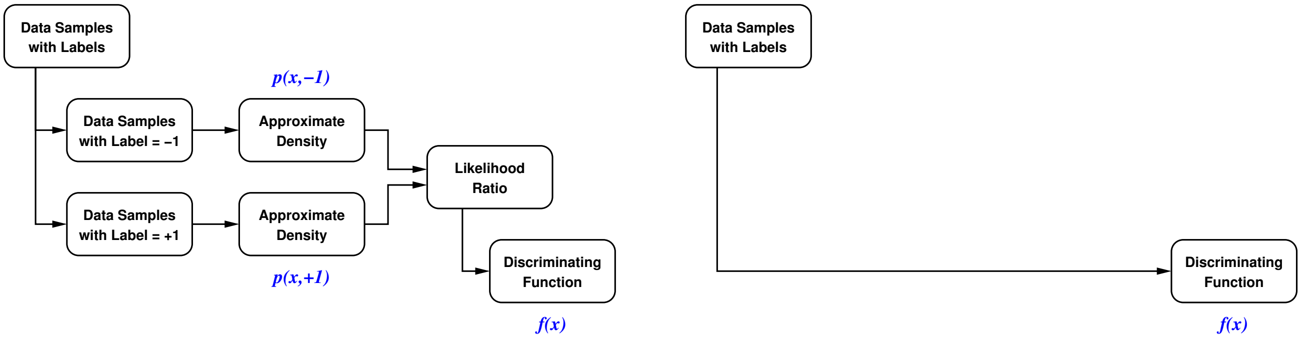
Email: jt@lanl.gov

If we *did* know $p(\mathbf{x}, y)$, then the likelihood ratio would give us an optimal detector

$$\mathcal{D}(\mathbf{x}) = \frac{p(\mathbf{x}, +1)}{p(\mathbf{x}, -1)} \tag{1}$$

and comparison with a threshold, $f(\mathbf{x}) = \text{sign}\,[\mathcal{D}(\mathbf{x}) - \theta]$, would provide the optimal classifier. The choice of threshold θ depends on two factors: one is the relative frequency of the two classes -1 and $+1$; and the other is the relative importance of the two types of errors: mistaking -1 for $+1$, versus mistaking $+1$ for -1. If both classes are equally frequent, and both errors are equally important, then the appropriate threshold is $\theta = 1$.

The likelihood ratio formulation suggests a conceptually straightforward two-step approach to binary classification. The first step is to estimate the densities $p(\mathbf{x}, -1)$ and $p(\mathbf{x}, +1)$ from the sample data, and the second step is to plug those estimates into Eq. (1). This is sometimes called a *generative* approach, because it estimates the model that was used to generate the data. Typically this is performed by assuming that the distribution $p(\mathbf{x}, y)$ has a given parametric form, and then estimating those parameters from the data samples; see Fig. 1(a). But Vapnik[2] advises against this: "When solving a given problem, try to avoid solving a more general problem as an intermediate step." Rather than estimate the density, which is a hard problem and is not what is ultimately needed, one should instead estimate $f(\mathbf{x})$ directly, as illustrated in Fig. 1(b). Algorithms that do this – and the support vector machine is a prime example[7] – lead to *discriminative* models.



(a) Generative modeling (b) Discriminative modeling

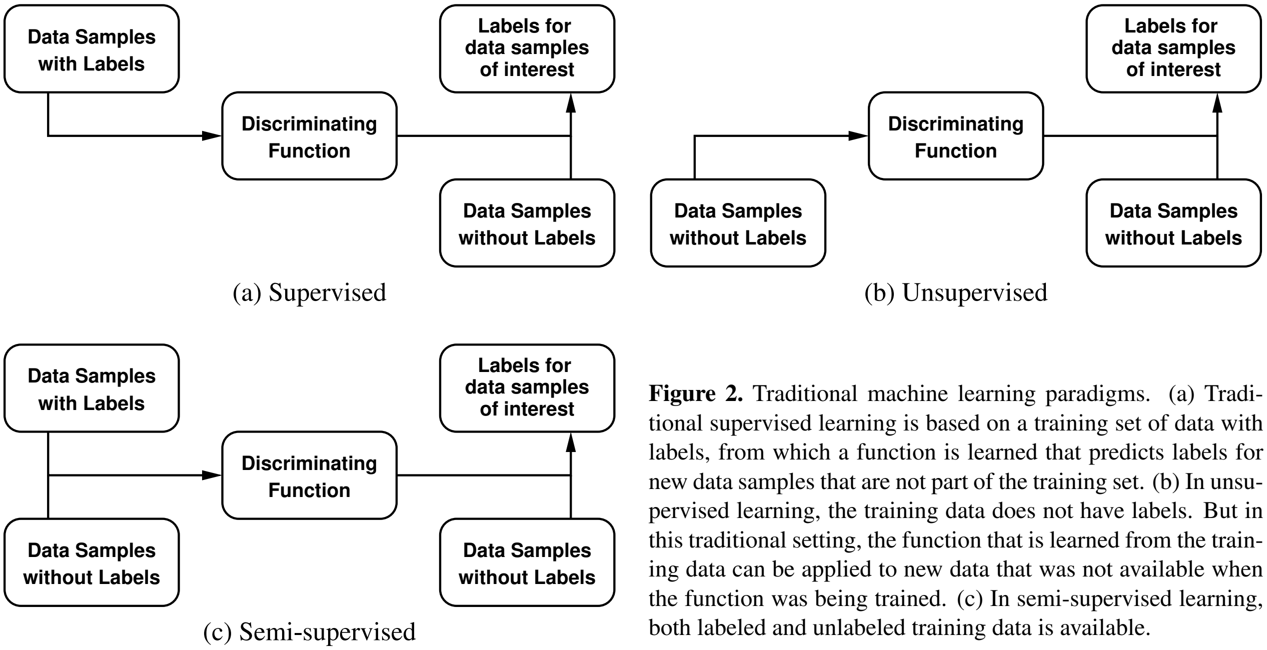
Figure 1. Although discriminative modeling has the *goal* as generative modeling, it is much less prescriptive about how to achieve it. That goal is the discriminating function $f(\mathbf{x})$, and the discriminative approach can avoid trying to solve a more general problem (*e.g.*, of approximating densities) as an intermediate step.

For this classic binary classification problem, the ultimate goal is the discriminating function $f(\mathbf{x})$, that can be applied to out-of-sample data. Since the training data is labeled, this is called supervised learning. An example of unsupervised classification is given by one-class classification, or anomaly detection. Here all of the training samples are assumed to come from a single class (for instance, the class of non-anomalous pixel spectra), and the discriminating function $f(\mathbf{x})$ can be used to predict whether new samples $\mathbf{x}$ belong to that class (if not, then they are considered anomalous). In semi-supervised learning, one has some data with labels and some data (usually a lot more of it) without labels, and the idea is to leverage the unlabeled data to improve the classifier $f(\mathbf{x})$ that would have been obtained using the (typically much less numerous) labeled data samples. The edited chapters in [8] describe a variety of approaches for taking advantage of auxiliary unlabeled data.

Fig. 2 illustrates these three kinds of "traditional" machine learning. In all three cases (supervised, unsupervised, and semi-supervised) the *inductive* learning is shown on the left, with in-sample data (labeled, unlabeled, or both) used to infer an approximating function. The *deductive* activity is shown on the right, where that function is applied to out-of-sample data.

3. TRANSDUCTIVE MACHINE LEARNING

The notion of transductive machine learning was introduced by Vapnik, and is illustrated in Fig. 3, which is adapted from Vapnik's book.[2] This figure lays out the philosophical premise of transductive learning, that it is a learning "from the particular to the particular," and is contrasted with inductive (from the particular to the general) and deductive (from the general to the particular). Vapnik argues that this is a valuable concept:


Data Samples with Labels
Discriminating Function
Labels for data samples of interest
Data Samples without Labels
Data Samples without Labels
Discriminating Function
Labels for data samples of interest
Data Samples without Labels
Data Samples with Labels
Discriminating Function
Labels for data samples of interest
Data Samples without Labels
Data Samples without Labels

(a) Supervised

(b) Unsupervised

(c) Semi-supervised

Figure 2. Traditional machine learning paradigms. (a) Traditional supervised learning is based on a training set of data with labels, from which a function is learned that predicts labels for new data samples that are not part of the training set. (b) In unsupervised learning, the training data does not have labels. But in this traditional setting, the function that is learned from the training data can be applied to new data that was not available when the function was being trained. (c) In semi-supervised learning, both labeled and unlabeled training data is available.

> The problem of estimating the values of a function at a given point addresses a question that has been discussed in philosophy for more than 2000 years: *What is the basis of human intelligence: knowledge of laws (rules) or the culture of direct access to the truth (intuition, adhoc inference)?* ...This model [transductive machine learning] can provide the strongest contribution to the 2000 years of discussions about the essence of human reason.[2]

In transductive learning, to be more specific, rather than treat the *general* function $f(\mathbf{x})$ as the final goal, we seek the *particular* function values $f(\mathbf{x}_i)$ for an unlabeled data set $\{\mathbf{x}_{n+1}, \ldots, \mathbf{x}_{n+n'}\}$ that is (in addition to the labeled data) available during training. This scenario was motivated by the principle of trying to avoid solving a more general problem as an intermediate step.

The distinction between in-sample and out-of-sample data has a different meaning in transductive learning. That unlabeled data set that is available during training, that is the data on which the algorithms is to be tested. Fig. 4 illustrates the transductive scenario a little bit more explicitly than Fig. 3. Comparing Fig. 4(a) to Fig. 2(a), the only difference is the role of the “Data Samples without Labels” box. For traditional learning, those samples are acted upon by the discriminating function; for transductive learning, those samples are used to help train that discriminating function, and *then* they are acted upon by the discriminating function. The usual taboo against testing on the training set doesn’t apply in this case, because those data samples are unlabeled.

For unsupervised learning, all of the data is unlabeled, and the distinction between traditional unsupervised and transductive unsupervised learning is drawn in Fig. 2(b) and Fig. 4(b). Most actual applications of unsupervised learning (*e.g.*, clustering or anomaly detection) usually work in the transductive scenario – *i.e.*, the algorithm is applied to the same data that the algorithm was trained on (in some applications, *e.g.*image segmentation, an explicit discimінating function may never have actually been produced).

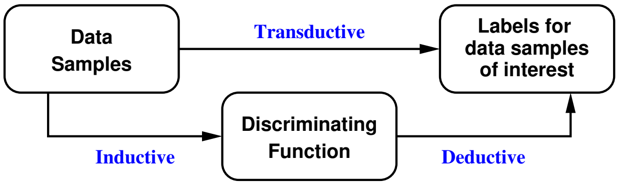
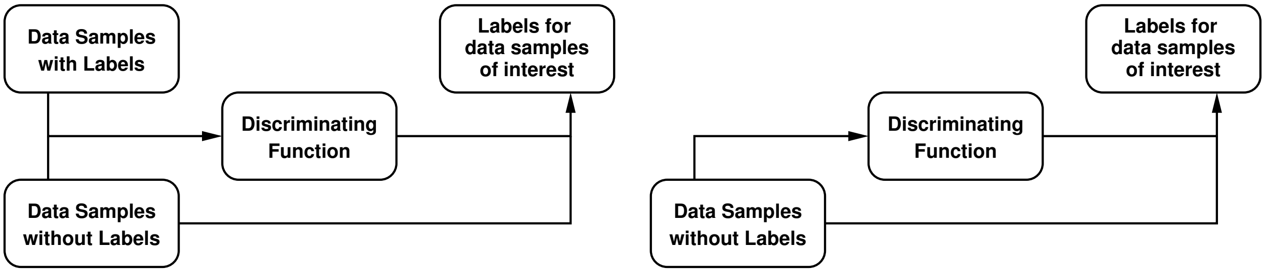


Figure 3. This figure, modified from Vapnik,[2] illustrates how transductive machine learning is related to inductive (specific to general) and deductive (general to specific) reasoning. Although this figure shows the transductive arrow bypassing the approximating function, virtually all practical transductive algorithms produce such a function as an intermediate step.


Data Samples with Labels
Discriminating Function
Labels for data samples of interest
Data Samples without Labels
Discriminating Function
Labels for data samples of interest
Data Samples without Labels

(a) Transductive

(b) Transductive/unsupervised

Figure 4. Transductive machine learning paradigms. (a) Transductive learning is similar to traditional semi-supervised learning, and the two are often confused. But in transductive learning, the final goal is not the function, but labels for the data that is available (but unlabeled) at the time of training. (b) Unsupervised transductive learning is similar to traditional unsupervised learning (and when unsupervised learning is done in practice, it is usually done transductively).

As a practical matter, transductive learning looks a lot like semi-supervised learning, and the two are often confused* but the distinction drawn here is best illustrated by comparing Fig. 2(c) to Fig. 4(a). The process of estimating the approximating function might be the same in both cases, but the use that is made of that function, and the evaluation of that function's performance, are what distinguish the two cases.

4. MATCHED-PAIR MACHINE LEARNING

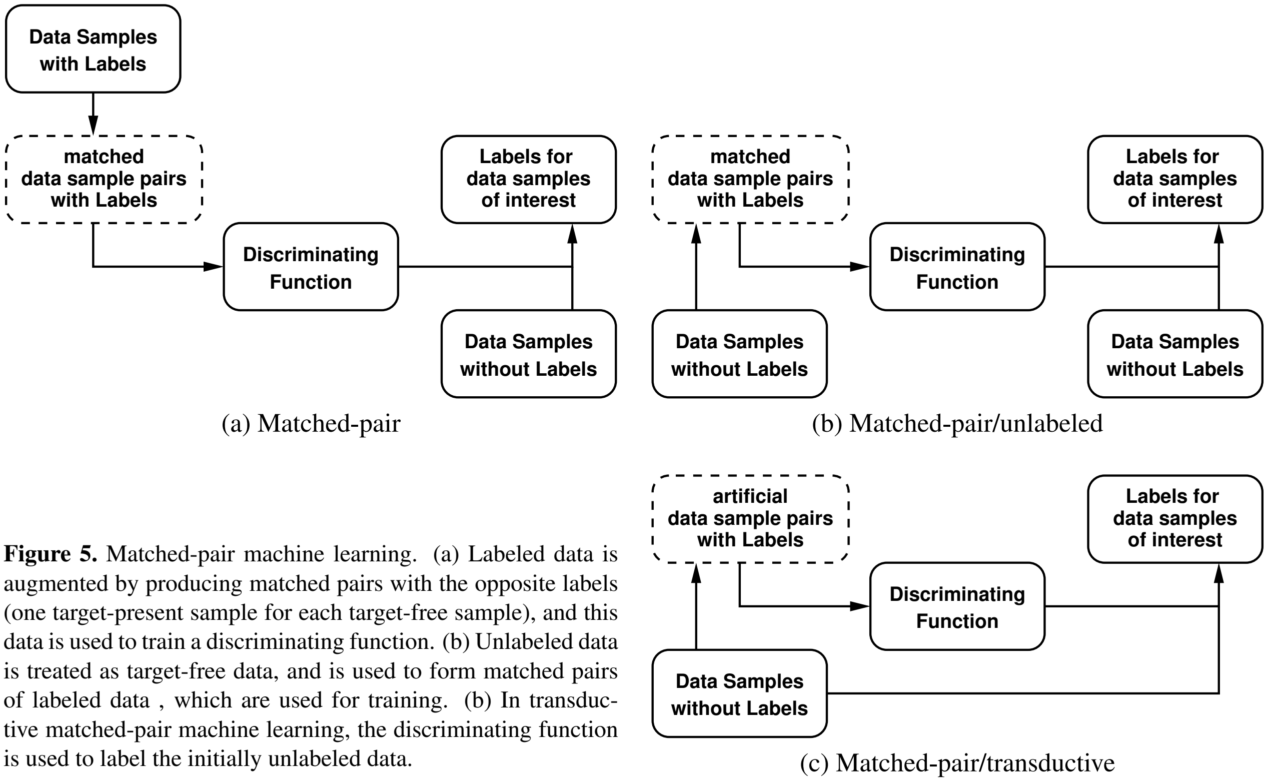
The term "matched pair" is taken from the statistics literature,[9, 10] and is used in hypothesis testing problems where the with-treatment and without-treatment samples are naturally paired. Taking that pairing into account leads to more sensitive tests of whether the given treatment has observable effects. The use of matched pairs in machine learning addresses a related, but different, problem. In matched-pair machine learning, the effect of a treatment is known, and encoded in a function $\boldsymbol{\xi}(\mathbf{x})$. The machine learning problem is to determine, for a given observation, whether one is observing $\mathbf{x}$ or $\boldsymbol{\xi}(\mathbf{x})$.

The reason this is so useful for target detection is that $\boldsymbol{\xi}(\mathbf{x})$ corresponds to what is sometimes called target implantation.[11, 12] For every non-target pixel whose spectrum is given by $\mathbf{x}$, an artificial target pixel can be created (or implanted) with a spectrum given by $\boldsymbol{\xi}(\mathbf{x})$. Although implantation can be used for *evaluating* detectors,[11, 12] the idea behind matched-pair machine learning is that implantation can also be used to create labeled training data for *learning* target detectors. The concept was employed by Foy *et al.*,[13, 14] as part of a larger exploration of the utility of a two-dimensional matched-filter-residual (MFR) projection space. A more systematic treatment of the matched-pair formalism was developed in [1].

As illustrated in Fig. 5, the matched-pair scheme works by replacing every sample $\mathbf{x}$, with a pair of samples: $\mathbf{x}$ and $\boldsymbol{\xi}(\mathbf{x})$; with $\mathbf{x}$ labeled non-target and $\boldsymbol{\xi}(\mathbf{x})$ labeled target. In the traditional formulation, shown in Fig. 5(a), the training is performed on data that is known not to include targets, and the discriminating function that is learned is then applied out-of-sample to new imagery that might include targets. If the training data, in addition to the preponderance of target-free samples, is contaminated with a few samples that do include targets, then the out-of-sample performance will suffer. But if the contaminating targets are sufficiently weak and/or rare, then the deleterious effect will be small.

A more practical scenario for target detection combines matched-pair learning with transductive learning. Here, one begins with unlabeled data (*e.g.*, pixels $\mathbf{x}_1, \ldots, \mathbf{x}_n$) that one presumed to be *mostly* target-free, but the goal is to find targets in that data. In the transductive matched-pair approach, one initially treats the data samples as if they were target-free, and creates labeled matched pairs: $(\mathbf{x}_i, -1)$, and $(\boldsymbol{\xi}(\mathbf{x}_i), +1)$. Using this as training data, one computes a discriminating function $f(\mathbf{x})$. In the traditional machine-learning scenario, the function $f(\mathbf{x})$ would then be applied to new data samples that have not yet been seen, but in the transductive scenario, $f(\mathbf{x})$ is applied back to the training data. This scenario effectively *assumes* that the training data are contaminated, and the goal is to find those contaminating samples, because those are the targets of interest.

*An entertaining essay on the topic appears as Chapter 25 (A Discussion of Semi-Supervised Learning and Transduction) of Ref. [8]. It is written as "a fictitious discussion inspired by real discussions between the editors of this book and a number of people, including Vladimir Vapnik."


Data Samples with Labels
matched data sample pairs with Labels
Discriminating Function
Labels for data samples of interest
Data Samples without Labels
matched data sample pairs with Labels
Discriminating Function
Labels for data samples of interest
Data Samples without Labels
Data Samples without Labels
artificial data sample pairs with Labels
Labels for data samples of interest
Discriminating Function
Data Samples without Labels

(a) Matched-pair

(b) Matched-pair/unlabeled

(c) Matched-pair/transductive

Figure 5. Matched-pair machine learning. (a) Labeled data is augmented by producing matched pairs with the opposite labels (one target-present sample for each target-free sample), and this data is used to train a discriminating function. (b) Unlabeled data is treated as target-free data, and is used to form matched pairs of labeled data , which are used for training. (b) In transductive matched-pair machine learning, the discriminating function is used to label the initially unlabeled data.

4.1. Examples of matched-pair opportunities

In [1], a number of non-remote-sensing applications were suggested (determination of white balance, detection of post-processing, identification of orientation, *i.e.*portrait *vs.* landscape, *etc.*) for which matched-pair learning might be effective. But the central example in [1] was the plume detection problem. Here, $\boldsymbol{\xi}(\mathbf{x})$ is the effect of plume on a background pixel $\mathbf{x}$. In particular, for an absorptive plume, Beer's law[15] provides $\boldsymbol{\xi}(\mathbf{x}) = \mathbf{x} \cdot \exp(-\epsilon\mathbf{t})$ where ϵ is a measure of plume strength and $\mathbf{t}$ is the spectrum of the chemical target of interest. For hyperspectral imaging in the thermal infrared, the effect of the plume can include both absorption and emission[16–21] but all of this can be included in $\boldsymbol{\xi}(\mathbf{x})$.

For subpixel target detection, using the "replacement model," one can write $\boldsymbol{\xi}(\mathbf{x}) = (1-\epsilon)\mathbf{x} + \epsilon\mathbf{t}$, where ϵ is the size (in fractions of a pixel) of the target, and $\mathbf{t}$ is its spectral signature. If the background distribution is Gaussian, then a generalized likelihood ratio test leads to the finite-target matched filter.[22] But for non-Gaussian backgrounds, one may prefer a detector that is adaptive to that background. In the MFR space, this was considered in [13, 14].

Shadows are an often confounding source of variability in remote sensing images, and a detector of shadows could be used to separate them from mainstream processing. If $\mathbf{x}$ is the pixel spectrum in sunlight, then one can describe a $\boldsymbol{\xi}(\mathbf{x})$ that corresponds to the spectrum of the same material in shadow. It will be darker, and in general, it will be bluer.

4.2. Nuisance parameters

Two of the examples in Section 4.1 involve a nuisance parameter corresponding to the strength of the target. This is generally not known in advance, and one seeks a detector that works over a range of possible strengths. Thus, this is not quite a binary classification problem, and the likelihood ratio in Eq. (1) does not quite apply. The customary approach is to consider a generalized likelihood ratio (GLR)

$$\mathcal{D}(\mathbf{x}) = \max_{\epsilon} \mathcal{D}(\mathbf{x};\epsilon), \quad \text{where} \tag{2}$$

$$\mathcal{D}(\mathbf{x};\epsilon) = \frac{p(\mathbf{x},+1;\epsilon)}{p(\mathbf{x},-1)}. \tag{3}$$

Although the non-optimality of the GLR has been widely documented,[23–26] it is nonetheless a straightforward and often successful approach to the composite hypothesis testing problem. A possible extension of the GLR to discriminative detectors is to apply Eq. (2) to a set of detectors that are derived not from a likelihood ratio, as in Eq. (3), but from a discriminative model (such as a support vector machine) trained on specific values $\epsilon_1, \epsilon_2, \ldots, \epsilon_m$. This is a kind of discrete clairvoyant fusion[27] that approximates continuum fusion.[24, 25]

An alternative, which has the advantage of always producing an "admissible" detector,[23] employs a Bayesian prior $p(\epsilon)$ on the nuisance variable. In this case (see also [28,29]), the detector is given by

$$\mathcal{D}(\mathbf{x}) = \int \mathcal{D}(\mathbf{x};\epsilon)p(\epsilon)\, d\epsilon \tag{4}$$

If one has an explicit expression for $\mathcal{D}(\mathbf{x};\epsilon)$ – which in turn requires explicit models for $p(\mathbf{x},-1)$ and $p(\mathbf{x},+1;\epsilon)$ – and a good choice for the prior $p(\epsilon)$, then the expression in Eq. (4) can be integrated to provide the detector $\mathcal{D}(\mathbf{x})$.

But if one wants to avoid modeling $p(\mathbf{x},-1)$ and $p(\mathbf{x},+1;\epsilon)$ as intermediate steps, then using the matched-pair formalism with a stochastic $\boldsymbol{\xi}(\mathbf{x})$ provides a way to obtain a discriminative model that shares with Eq. (4) the property of admissibility. This is illustrated in the next section.

4.3. Illustrative Example with Simulated Data

Consider the subpixel target problem with a replacement model: $\boldsymbol{\xi}(\mathbf{x}) = (1-\epsilon)\mathbf{x} + \epsilon\mathbf{t}$. Here $0 \le \epsilon \le 1$ is the fraction of a pixel covered by the target, and in general it is not known. Our approach here, consistent with Eq. (4), is to treat it as a random variable. Thus, for every data sample $\mathbf{x}_i$, a matched pair is created:

$$(\mathbf{x}_i, -1), \tag{5}$$

$$((1-\epsilon_i)\mathbf{x}_i + \epsilon_i\mathbf{t}, +1), \tag{6}$$

where ϵ_i is randomly chosen from a uniform distribution over the range $[0, 1]$.

In the experiment, whose results are shown in Fig. 6, the background $\mathbf{x}$ is drawn from a $d = 21$ dimensional elliptically-contoured multivariate-t distribution,[30] using parameter $\nu = \infty$ (Gaussian) or $\nu = 5$ (heavy-tailed). The use of different background models emphasizes the importance of background modeling in target detection.[31] The target signature $\mathbf{t}$ is known, and is taken here to be of magnitude $\|\mathbf{t}\| = 5$. Using the matched-pair data for training, both a linear adaptive matched filter and a support vector machine (SVM) classifier (as implemented in the python package *scikit-learn*[32]) are employed. These detectors are applied in the matched-filter-residual (MFR) space defined in [13, 14]. In MFR space, the vertical axis corresponds to the matched filter direction, which in this case (since the covariance of the clutter is white) corresponds to the direction $\hat{\mathbf{t}} = \mathbf{t}/\|\mathbf{t}\|$; the horizontal axis corresponds to the magnitude of the residual. Thus:

$$\text{MF} = \hat{\mathbf{t}}^T\mathbf{x}, \tag{7}$$

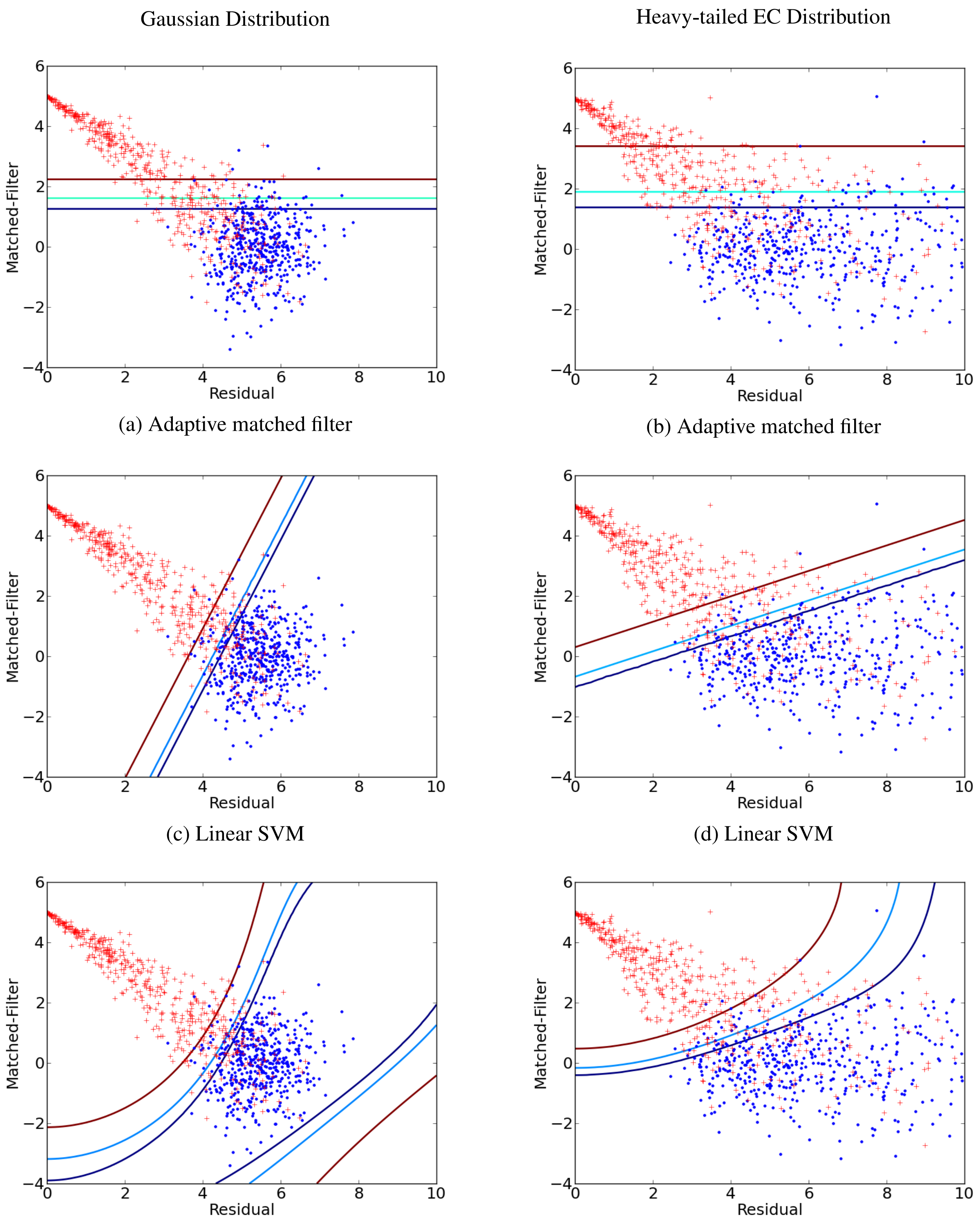
$$\text{R} = \|(I - \hat{\mathbf{t}}\hat{\mathbf{t}}^T)\mathbf{x}\| = \sqrt{\|\mathbf{x}\|^2 - (\text{MF})^2}. \tag{8}$$

Three different detectors were used for this experiment.

Fig. 6(a,b) shows the simple adaptive matched filter,[33, 34] whose contours are horizontal lines since the vertical axis corresponds to the matched filter output. Visual inspection of the on-target and off-target points on those plots shows poor separation, a result that confirms what various authors have already reported: namely that the additive model is inappropriate for subpixel target detection.[22, 35, 36]

Fig. 6(c,d) shows a nonlinear detector that is nonetheless linear in the (nonlinear) MFR space. It is computed as a linear SVM and it provides much better separation of on-target and off-target pixels than does the matched filter in Fig. 6(a,b). Comparing Fig. 6(c) to Fig. 6(d) also shows the importance of adapting the detector to the background distribution. It is interesting to note, in Fig. 6(d), that the intercept of the linear detector is near zero. An intercept that is strictly zero would correspond to the adaptive coherence estimation (ACE) detector.[37, 38]

Finally, in Fig. 6(e,f), the full flexibility of the MFR space is employed, and nonlinear detectors are derived using the kernelized SVM. A classic radial basis function kernel is used: $k(\mathbf{x}_1, \mathbf{x}_2) = \exp(-\gamma\|\mathbf{x}_1 - \mathbf{x}_2\|^2)$, where a small $\gamma = 0.03$ was chosen to avoid over-fitting. Again, we observe good separation of on-target and off-target pixels (better than the simple matched filter, anyway). It may be noteworthy that the contours in Fig. 6(f) are reminiscent of the EC-GLRT solution reported in [39].



(a) Adaptive matched filter

(b) Adaptive matched filter

(c) Linear SVM

(d) Linear SVM

(e) Radial-basis SVM

(f) Radial-basis SVM

Figure 6. Illustrative example with simulated data, shown in matched-filter versus residual (MFR) space. Blue dots are the target-free $\mathbf{x}$ values, which are drawn from an elliptically contoured multivariate distribution that is either Gaussian (a,c,e) or heavy-tailed (b,d,f). Red plus symbols are the target-present values, given by $\boldsymbol{\xi}(\mathbf{x}) = (1 - \epsilon)\mathbf{x} + \epsilon\mathbf{t}$, with ϵ chosen at random in the range $[0, 1]$. For all of the detectors shown, the three contours correspond to false alarm rates of one percent, five percent, and ten percent.

5. CONCLUSION

Matched-pair machine learning is well suited to the target detection problem because it enables an approach that allows physical modeling of the target to be incorporated into data-driven modeling of the background. This addresses the shortcoming of purely data-driven machine learning that the number of target-class samples is usually too small to enable representative statistical characterization of targets. It also addresses the shortcoming of direct modeling approaches that they are too inflexible to account for the complexity of observed hyperspectral clutter.

The transductive scenario is also well suited to target detection because it provides a hedge against non-stationarity in the background by using a background model that is effectively "trained" on the same data to which it is applied. The transductive approach furthermore works well in combination with the matched-pair approach for practical target detection scenarios.

ACKNOWLEDGMENTS

I am grateful to Don Hush and Reid Porter for enlightening discussions on the topic of transductive reasoning, and I further thank Reid for pointing me to the chapter in Ref. [8] on the difference between semi-supervised and transductive learning. I am also grateful to Bernie Foy for early discussions of matched-pair learning in the context of target detection. This work was supported by the Laboratory Directed Research and Development (LDRD) program at Los Alamos National Laboratory.

REFERENCES

1. J. Theiler, "Matched-pair machine learning," *Technometrics* **55**, pp. 536–547, 2013.
2. V. Vapnik, *The Nature of Statistical Learning Theory*, Springer, New York, 2nd ed., 1999.
3. J. A. Gualtieri and R. F. Cromp, "Support vector machines for hyperspectral remote sensing classification," *Proc. SPIE* **3584**, pp. 221–232, 1999.
4. S. J. Perkins, N. R. Harvey, S. P. Brumby, and K. Lacker, "Support vector machines for broad-area feature classification in remotely sensed images," *Proc. SPIE* **4381**, pp. 286–295, 2001.
5. N. R. Harvey, J. Theiler, S. P. Brumby, S. Perkins, J. J. Szymanski, J. J. Bloch, R. B. Porter, M. Galassi, and A. C. Young, "Comparison of GENIE and conventional supervised classifiers for multispectral image feature extraction," *IEEE Transactions on Geoscience and Remote Sensing* **40**, pp. 393–404, 2002.
6. G. Camps-Valls, D. Tuia, L. Bruzzone, and J. A. Benediktsson, "Advances in hyperspectral image classification," *Signal Processing Magazine* **31**, pp. 45–54, January 2014.
7. C. Cortes and V. Vapnik, "Support vector machine," *Machine Learning* **20**, pp. 273–297, 1995.
8. O. Chapelle, B. Schölkopf, and A. Zien, *Semi-Supervised Learning*, MIT Press, Cambridge, MA, 2006.
9. D. B. Rubin, "Matching to remove bias in observational studies," *Biometrics* **29**, pp. 159–183, 1973.
10. M. J. Crawley, *Statistics: an introduction using R*, Wiley, West Sussex, UK, 2005.
11. W. F. Basener, E. Nance, and J. Kerekes, "The target implant method for predicting target difficulty and detector performance in hyperspectral imagery," *Proc. SPIE* **8048**, p. 80481H, 2011.
12. Y. Cohen, Y. August, D. G. Blumberg, and S. R. Rotman, "Evaluating sub-pixel target detection algorithms in hyperspectral imagery," *J. Electrical and Computer Engineering* **2012**, p. 103286, 2012.
13. B. R. Foy, J. Theiler, and A. M. Fraser, "Hyperspectral target detection using machine learning in a reduced dimensional space," *Proc. MSS (Military Sensing Sympoisia) Passive Sensors Conference* , 2008.
14. B. R. Foy, J. Theiler, and A. M. Fraser, "Decision boundaries in two dimensions for target detection in hyperspectral imagery," *Optics Express* **17**, pp. 17391–17411, 2009.
15. A. Beer, "Bestimmung der absorption des rothen lichts in farbigen flussigketiten," *Ann. Physik* **86**, pp. 78–88, 1852.
16. C. L. Bennett, M. R. Carter, D. J. Fields, and F. D. Lee, "Infrared hyperspectral imaging results from vapor plume experiments," *Proc. SPIE* **2480**, pp. 435–444, 1995.
17. A. Hayden, E. Niple, and B. Boyce, "Determination of trace-gas amounts in plumes by the use of orthogonal digital filtering of thermal-emission spectra," *Applied Optics* **35**, pp. 2802–2809, 1996.
18. S. J. Young, "Detection and quantification of gases in industrial-stack plumes using thermal-infrared hyperspectral imaging," Tech. Rep. ATR-2002(8407)-1, The Aerospace Corporation, 2002.

19. B. R. Foy, H. A. Fry, and B. D. McVey, "Approaches to chemical plume detection in hyperspectral infrared imaging," Tech. Rep. LA-CP-03-0103, Los Alamos National Laboratory, 2003.
20. J. Theiler, B. R. Foy, and A. M. Fraser, "Characterizing non-Gaussian clutter and detecting weak gaseous plumes in hyperspectral imagery," *Proc. SPIE* **5806**, pp. 182–193, 2005.
21. C. M. Gittins, "Detection and characterization of chemical vapor fugitive emissions by nonlinear optimal estimation: theory and simulation," *Applied Optics* **48**, pp. 4545–4561, 2009.
22. A. Schaum and A. Stocker, "Spectrally selective target detection," *Proc. ISSSR (International Symposium on Spectral Sensing Research)* , p. 23, 1997.
23. E. L. Lehmann and J. P. Romano, *Testing Statistical Hypotheses*, Springer, New York, 2005.
24. A. Schaum, "Continuum fusion: a theory of inference, with applications to hyperspectral detection," *Optics Express* **18**, pp. 8171–8181, 2010.
25. A. P. Schaum and B. J. Daniel, "Continuum fusion methods of spectral detection," *Optical Engineering* **51**, p. 111718, 2012.
26. J. Theiler, "Confusion and clairvoyance: some remarks on the composite hypothesis testing problem," *Proc. SPIE* **8390**, p. 839003, 2012.
27. J. Theiler, "Formulation for min-max clairvoyant fusion based on monotonic recalibration of statistics," *Optical Engineering* **51**, p. 111714, 2012.
28. A. Schaum, "Hyperspectral target detection using a Bayesian likelihood ratio test," *Proc. IEEE Aerospace Conference* **3**, pp. 1537–1540, 2002.
29. J. Theiler, B. R. Foy, and A. M. Fraser, "Beyond the adaptive matched filter: nonlinear detectors for weak signals in high-dimensional clutter," *Proc. SPIE* **6565**, p. 656503, 2007.
30. D. Manolakis, D. Marden, J. Kerekes, and G. Shaw, "On the statistics of hyperspectral imaging data," *Proc. SPIE* **4381**, pp. 308–316, 2001.
31. S. Matteoli, M. Diani, and J. Theiler, "An overview background modeling for detection of targets and anomalies in hyperspectral remotely sensed imagery," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (JSTARS)* **7**, 2014. To appear (doi:10.1109/JSTARS.2014.2315772).
32. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research* **12**, pp. 2825–2830, 2011.
33. I. S. Reed, J. D. Mallett, and L. E. Brennan, "Rapid convergence rate in adaptive arrays," *IEEE Trans. Aerospace and Electronic Systems* **10**, pp. 853–863, 1974.
34. F. C. Robey, D. R. Fuhrmann, E. J. Kelly, and R. Nitzberg, "A CFAR adaptive matched filter detector," *IEEE Trans. Aerospace and Electronic Systems* **28**, pp. 208–216, 1992.
35. D. Manolakis, E. Truslow, M. Pieper, T. Cooley, and M. Brueggeman, "Detection algorithms in hyperspectral imaging systems: An overview of practical algorithms," *IEEE Signal Processing Magazine* **31**, pp. 24 – 33, Jan 2014.
36. A. Schaum, "Continuum fusion solutions for replacement target models in electro-optic detection," *Applied Optics* **53**, pp. C25–C31, 2014.
37. L. L. Scharf and L. T. McWhorter, "Adaptive matched subspace detectors and adaptive coherence estimators," *Proc. 30th Asilomar Conference on Signals, Systems, and Computers* , pp. 1114–1117, 1996.
38. S. Kraut, L. L. Scharf, and R. W. Butler, "The Adaptive Coherence Estimator: a uniformly most-powerful-invariant adaptive detection statistic," *IEEE Trans. Signal Processing* **53**, pp. 427–438, 2005.
39. J. Theiler and B. R. Foy, "EC-GLRT: Detecting weak plumes in non-Gaussian hyperspectral clutter using an elliptically-contoured generalized likelihood ratio test," *Proc. IEEE International Geoscience and Remote Sensing Symposium (IGARSS)* , p. I:221, 2008.